CURSE OF DIMENSIONALITY

→ When is Data high Dimensional and Why Might That be a problem?

When data has huge number of dimensions :

1. If we have more features than observations than we run the risk of massively overfitting our model - this would generally result in terrible out of sample performance.
2. When we have too many features, observations become harder to cluster. Too many dimensions causes every observation in dataset to appear equidistant from all the others. And because clustering uses Euclidean distance to quantify similary distance between observations then all distance become approximately equal / alike & no meaningful cluster formed.

Example - Let take 8 dishes (food) - 1. Rosgulla. 3. Sonpapdi 5. Panner tikka 7. Spicy Veg Mix 8. Mutton curry 2. Tandoori Chicken 4. Eclairs 6. 5star

Instead of using 8 variables, we can use two clusters - 1. Sweet 2. Spicy

Now Sweet → 1) Rosgulla 2) Eclairs 3) 5 star 4) Sonpapdi.

Spicy → 1) Tandoori chicken 2) Panner tikka 3) Mutton curry 4) Spicy veg mix

But to differentiate into two dimension, is actually not simple. But a machine learning algorithim can do if data is presented properly.

→ So if instead of 2 categories, we have 8 then the classification will be difficult in the testing set because

i) Every dish have own specification

ii) As an algorithim I don't know relationship between dishes

→ But if we do feature reduction I know, 4 are spicy and sweet respectively

So Dimensionality Reduction methods are → i) Principal Component Analysis (PCA)

ii) Kernel PCA

iii) Linear Discriminant Analysis (LDA)